SAN DIEGO STATE UNIVERSITY

Antoni Luque Santolaria <aluque@sdsu.edu>

More "super-short" Caudovirales data

Stephen Nayfach <snayfach@lbl.gov> Thu, Dec 10, 2020 at 6:02 PM
To: Antoni Luque Santolaria <aluque@sdsu.edu>
Cc: Simon Roux <sroux@lbl.gov>, "Nikos C. Kyrpides" <nckyrpides@lbl.gov>, sean benler <sbenler@gmail.com>, "White, Simon" <simon.white@uconn.edu>

Hi Toni,

This sounds super interesting!

Here's a dropbox link to 7,650 viral genomes with the following properties:
-less than 20kb
-are predicted to be circular (contain a direct terminal repeat of at least 20-bp)
-contain at least 1 Caudovirales marker gene
-were derived from a number of diverse datasets, but mostly bulk metagenomes w/o viral enrichment

Most of the fields in the TSV file should be self-explanatory, but please let me know if you have any questions or if you have issues with the link. The field "gca_name" indicates whether a sequence is similar to a GenBank genome.

We expect many/most of these sequences to be genomic fragments, but it would be cool to find some putatively complete Caudovirales with unusually small capsids.

Interested to see what you find :)

Best,
Stephen

On Thu, Dec 10, 2020 at 7:21 AM Antoni Luque Santolaria <aluque@sdsu.edu> wrote:

> Hi Simon and Stephen,
>
> It would be great to scrutinizing the ~4,000 circular contigs with < 10kb. We only found 17 below 10 kb in our analysis of gut metagenomes. The model predicts that there could be tailed phages as small as ~1.3 kbp on the lowest limit, so there is plenty of room there for potential and unusual tailed phages that have not been characterized (see Figure 4 in our paper---attached as png). The range 10–17 kb is also relevant. There are a couple of capsid architectures missing there. We predicted some isolates to adopt one of them (T=3) (see Figure 5 in our paper---also attached as png).
>
> Some of the smallest circular contigs that we found in gut metagenome-assembled genomes did not contain MCPs or a portal (see Supplementary Figure S3). In the paper, we discussed that they could be phage satellites. I'm adding Sean Benler in CC (from Eugene Koonin's lab). He reconstructed and analyzed the circular genomes from gut metagenomes in our paper and in another recent paper (Benler et al: ref 64). We are now scrutinizing the small genomes with more stringent criteria. We are uncertain about those in the 5 kb range, but we are very confident with most contigs 7.5 kb or larger, which, in any case, gives us room for new tailed phage structures.

I'm adding Simon White in CC too. He is an assistant professor in structural biology at the University of Connecticut and co-author of the paper. We are collaborating to characterize tailed phage structures. Investigating small tailed phages was an idea related to the projects that we are developing together. The CheckV tool would align perfectly with our goal to sample the existence of uncharacterized capsid structures in the environment. Investigating contigs 10 kb or smaller and their distribution would be priceless.

If possible, I would like to provide predictions of capsid structures for the circular genomes obtained in CheckV and analyze in-depth the frequency and ecosystem distribution of < 20 kb contigs.

Let me know if any of you would be interested in collaborating to investigating this.

Thanks,
Toni

On Wed, Dec 9, 2020 at 4:11 PM Simon Roux <sroux@lbl.gov> wrote:

> Hi Toni,
>
> Following up on this question of finding short Caudovirales genomes, I checked with Stephen Nayfach (a scientist in the group of Nikos Kyrpides at JGI), who may have more candidates than are currently shown into IMG. (Nikos, for context, Toni recently published a paper that predicted the existence of a set of small tailed phage based on the geometry of the capsids, which should have very short genomes, probably < 10kb: https://www.mdpi.com/2076-2607/8/12/1944). Stephen recently performed a large search of all potential complete viral genomes from metagenomes (for a tool he developed called CheckV). During this search, he found about ~ 4,000 contigs which had direct terminal repeats (i.e. "circular"), but were discarded because they were shorter than 10kb. We assumed these were assembly artifacts, and certainly a number of them are, but these may also include some of these short Caudovirales you predicted should exist ?
>
> I've CC'd Stephen and Nikos here, so that you guys can interact directly if this is an avenue you would like to pursue.
>
> Best,
> Simon

--
Dr. Antoni Luque
http://luquelab.com
Google Scholar profile
Assistant Professor
Department of Mathematics and Statistics
Viral Information Institute
Computational Science Research Center
San Diego State University